AD_____

Award Number: DAMD17-02-1-0499

TITLE: Cloning and Characterization of Expanded CAG-Repeat Containing Sequence(s): Identification of Candidate Breast Cancer Predisposition Gene(s)

PRINCIPAL INVESTIGATOR: Hilmi Ozcelik, Ph.D.
Hamdi Jarjanazi
Noel Pabalan

CONTRACTING ORGANIZATION: Mount Sinai Hospital
Toronto, Ontario M5G 1X5
Canada

REPORT DATE: June 2005

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-06-2005	Final	1 May 2002 - 30 May 2005

4. TITLE AND SUBTITLE
Cloning and Characterization of Expanded CAG-Repeat Containing Sequence(s): Identification of Candidate Breast Cancer Predisposition Gene(s)

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-02-1-0499

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Hilmi Ozcelik, Ph.D.; Hamdi Jarjanazi; Noel Pabalan;

E-mail: ozcelik@mshri.on.ca

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Mount Sinai Hospital
Toronto, Ontario M5G 1X5
Canada

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Trinucleotide repeats (TNRs) are widely present in the human genome, and their expansions have been recognized to be the cause of several genetic disorders. In a previous study, we have identified expanded (CAG) repeats in 2.4% of breast cancer cases (n=212). No expansion of the same magnitude has been detected in 196 population controls. In the current project our objective is to identify and characterize such sequences in breast cancer specimens. We have developed an effective cloning and screening strategy using a combination of dynabead enrichment microsatellite isolation protocol (Glenn-Schable protocol). Colonies were selected and sequenced to identify CAG repeat containing fragments from breast cancer patients. We have shown that ~20% of the colonies contained polymorphic (larger/smaller) repeat lengths compared to corresponding sequences deposited at NCBI. Such sequences resemble polymorphic CAG repeats that may influence the intrinsic properties and function of nearby genes. To complement the library construction, we have also systematically identified CAG repeat containing coding sequences from cancer genes using Bioinformatics. In this study we have implicated the presence of polymorphic CAG repeats within the functional domains of cancer proteins. Molecular cloning and bioinformatics approaches used in this study has provided a valuable resource for polymorphic CAG-containing sequences within relevant cancer genes. Future direction involves genetic and epidemiological studies, where the risk contributed by CAG repeat expansions to breast cancer risk will be evaluated.

15. SUBJECT TERMS
Breast cancer, predisposition, trinucleotide, CAG repeats, cloning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	42	

I. INTRODUCTION:

Expression of an inherited disease gene does not always follow a Mendelian inheritance pattern. Since the turn of the century, it has been observed that the transmission of certain disease genes from one generation to the next is associated with an increase in severity of the disease symptoms and/or a decrease in the age of onset. This clinical phenomenon is called anticipation and was observed with many inherited diseases, especially the ones involving neurological disorders. However, the absence of a satisfactory explanation for anticipation at the molecular level resulted in much controversy in the past over acceptance of this concept.

Advances in molecular genetic analyses of inherited human diseases in which anticipation has been observed has recently attributed this phenomenon to a new class of dynamic mutations, characterized by trinucleotide repeat expansions (TREs) in loci where disease genes have been mapped.

The influence of TREs to cellular function depends on the location of the repeats relative to the gene and the type of repeats (for review, see Sanjeeva and Housman, 1997, Margolis et al., 1999, Vincent et al., 2000). The molecular consequences of TREs and the mechanism by which they result in the pathological condition may be quite diverse. In general however, TREs have been shown to perturb either structure and function (type I mutations) or expression (type II mutations) of the affected gene (for review see Sanjeeva and Housman, 1997, Margolis et al., 1999, Vincent et al., 2000). For example, expansion of a (CAG)n-repeat in the coding region of the gene for Huntington's disease results in the expression of an altered protein which contains an expanded polyglutamine region. This in turn leads to altered conformation, processing and general physical properties of the protein function of the product (Trottier et al., 1995).

Alternatively, expansions that result in 200-2000 CGG repeats in the 5' untranslated region of the Fragile X syndrome (FRAXA) gene results in loss of expression of FRAXA mRNA (Pieretti et. al., 1991). As the repeat expands with transmission to the next generation, the CGG repeats become more methylated reducing transcription of the FRAXA gene. The nature of this dynamic group of mutations may involve very large amplification of trinucleotide repeats (TNRs) rendering the sequence unstable during meiosis and resulting in intergenerational instability of the length of the TNR. It is not known why repeats that exceed a critical value are unstably transmitted to succeeding generations with a tendency towards expansion of TNRs.

There is however, an unambiguous association between longer expansions at the disease loci in succeeding generations and earlier clinical manifestation (Sanjeeva and Housman, 1997, Margolis et al., 1999, Vincent et al., 2000). Anticipation, therefore, is now commonly accepted as a hallmark of inheritance of an amplified TRE mutation. Thus far, at least 12 genetic disorders (mostly muscular or neurological) have been attributed to expansions of TNRs in loci containing the disease gene. These diseases are characterized as having increasing copy numbers of unstable expanded sequences in subsequent generations.

One can envision that expansions of TNRs are not restricted to neuro-muscular disorders, and they probably represent a novel class of dynamic mutations causing various human diseases. A study by O'Donovan et al. (1996), provided an intriguing finding suggesting that repeat expansions have a widespread role in common human

diseases. This study has shown that older, healthy individuals have generally shorter CAG-repeat lengths than their younger counterparts. The demonstration of increased, genome wide repeat-copy number with declining age in healthy populations suggests that dynamic mutation may have an encompassing role in human susceptibility to disease and that TNR diseases do not display a single major gene inheritance pattern.

II. STATEMENT OF WORK

Technical Objective 1: Cloning of Gene Sequences with CAG-Repeat Expansion.
(Four parallel cloning procedures will be performed simultaneously using 4 pools each including DNA mixed from 2 RED positive cases)

Task 1: Months 1-4	Enrichment by digestion of DNA samples and separation on agarose gels. Slicing (2-4mm), and extracting DNA from each slice (approximately 200-400 slices). Apply RED analysis on all slices to locate the fragments with expansion.
Task 2: Months 4-6	Cloning of the DNA into a ZapII system, electroporation into *E.Coli*, and amplification of the bacteria. Secondary enrichment by pooling: Plating the bacteria, pooling and amplification of pools. Extraction of DNA (~100-200) and RED analysis (~ 100-200).
Task 3: Months 6-8	Transformation of RED positive DNA into *E.Coli* CJ236 strain and generation of ssDNA. Isolation of ssDNA and production of dsDNA containing only CAG inserts CAG-probes and the primer extension method. Electroporation into *E.Coli*, amplification, and plating. Selection and amplification of individual clones(~100-200) and subsequent extraction of DNA.
Task 4: Months 8-10	RED analysis of the DNA extracted from individual clones (~100-200). Sequencing of inserts from RED positive clones. Identification of clones with large repeat sequences.
Task 5: Months 10-11	Designing PCR primers for every sequence with large CAG-repeats identified through cloning. Screening of all the cases originally detected to have the CAG-repeat expansion (detected by RED) by PCR and sequencing.

Technical Objective 2: Identification and Characterization of Gene(s) containing or Flanking Expanded CAG-Repeats

Task 1: Months 11-12	Designing PCR protocols for sequences that have large CAG-repeats through cloning and optimization of their detection. Sequencing of a small panel of control specimens to identify repeats with varying sizes to be used as size controls on microsatellite gels.

Task 2: Months 12-17 — Microsatellite analysis of 100 control specimens for every repeat cloned to determine allelic frequency of each repeat. It is expected than several regions containing large CAG-repeats will be identified from cloning of RED positive DNA from 8 breast cancer cases.

Task 3: Months 17-22 — Extraction of sequence data using different sequence database sources including GenBank, EST and STS. Synthesizing disparate pieces of information to find longer sequence data (or patterns). Mapping of CAG repeats using public genetic map resources, and identifying other genes around the CAG-repeat. Locating exact positions of CAG repeats relative to the structure of genes identified.

Task 4: Months 20-24 — Obtaining sequence data from databases and interpreting this information utilizing extensive literature searches. If necessary, wet lab work will be carried out to clarify the position of CAG-repeat in relation to the structure of gene(s) identified. Searches for information published on the loci/genes found and their associations with breast cancer. Writing of manuscripts for submission to peer reviewed journals.

III. BODY

III.A. Overview

Our goal is to clone and characterize expanded CAG repeat containing sequences in breast cancer patients. Such expanded sequences may influence the function of potential breast cancer predisposition genes. Most of the reported methods utilized for library construction and cloning of large repeat containing fragments require the use of a relatively large amount of genomic DNA as starting template (Yuan et al, 2001; Vincent et al, 2000; Koob et al, 1998). Given that patient material is a limited resource, it was to our interest in developing an efficient cloning strategy for enrichment with minimum use of starting genomic DNA. The tendency of long TNRs form special secondary structures reduces the efficiency of cloning fragments and makes the method cumbersome (Koob et al, 1998, Sanpei et al, 1996). Isolation procedure of TNRs from genomic DNA has been established and efficiency of the method has been evaluated in the previous report.

The protocol, which was modified from Inoue et al (1999) and Tozaki et al (2000) is restricted to three nucleotide substrates and primed by a biotinylated probe. Sequences containing TNRs are then isolated by a streptavidine biotin trapping method. Using this method, we had successfully detected the repeat containing fragments, albeit at a relatively low efficiency (60%). More recently, we applied the microsatellite isolation with Dynabeads protocol of Travis Glenn and Mandy Schable at the Department of Biological Sciences, University of South Carolina, a refined version from the method described by Hamilton et al. (1999). This procedure was found to be more efficient for our project because it increased the percentage of identifying repeat containing

fragments up to 80%. We modified the Glenn-Schable protocol by increasing the stringency of the method in an effort to reduce formation of secondary structures in repeat containing fragments.

III.B. Progress

B.1. Development of Repeat-Sequence Cloning Methodology

B.1.a. Various Cloning Methods Evaluated

Isolation of TNRs from breast cancer patient samples involved application of two protocols. These are the Tozaki et al (2000) and Glenn-Schable (2003) methods. Both involve enrichments that included hybridization-capture of repeat regions using biotinylated oligonucleotides with subsequent amplification. The Tozaki et al method produced sequences with relatively low percentage of repeats while the Glenn-Schable protocol produced relatively high percentage of sequences with TNRs.

B.1.b. Selection of the Glenn-Schable Protocol

The Glenn-Schable protocol is actually a refined synthesis of previous microsatellite (MS) isolation methods (e.g. Hamilton et al. 1999) that is characterized by the speed and ease with which multiple samples are processed. This method relied on cloning kits that invariably speeded up the process of TNR isolation. Moreover, efficiency of capture of strands with repeats is increased with two rounds of enrichment with biotinylated oligonucleotides.

B.1.c. Establishment of the Glenn-Schable Protocol

Genomic DNA was fragmented using multiple enzymes (Rsa I and BsTU I). The DNA strands were prepared for amplification by ligating a linker (Super SNX) to each DNA fragment, which provided the primer-binding site for the subsequent PCR step. Dynabead enrichment captured amplicons with repeat sequences and washed away all other fragments. The DNA containing repeats were then amplified. Next, two types of incorporation were applied to the amplicons using Invitrogen's TA cloning kits. The first was insertion of the DNA fragments into a cloning vector. The second involved transformation of the vector-insert by incorporation into a bacterial host. The bacteria were plated onto agar and incubated for 12-36h at 37°C. From these plates, isolated colonies were picked and dipped into LB broth, which was incubated for 6h at 37°C. DNA in the broth was used for PCR, the product of which was purified with exonuclease (EXO I) and shrimp alkaline phosphatase (SAP) prior to sequencing. The resulting DNA sequences were analysed *in silico*. This involved generation of a local database that included all relevant information about the sequence of each sample.

Optimizations are akin to detours along the route to isolating TNRs in order to determine the best conditions for amplification of DNA with repeat sequences. This involved mainly titrations of annealing temperatures (65°C and 72°C) and inclusion of denaturing agents (DMSO and glycerol) subsequent to the initial enrichment step. This was done in an effort to reduce formation of secondary structures in the sequence products.

B.1.d. Comparison to Other Existing Techniques

Other techniques used to isolate MS include those of Koob *et al* (1998) and Yuan *et al* (2001). The method used by Koob *et al* involves isolation of expanded TNRs and corresponding flanking sequences. It is accomplished through a process called RAPID (Repeat Analysis Pooled Isolation and Detection) cloning in which one, isolated clone is obtained. The approach of Yuan *et al* is based on size-separation of genomic fragments followed by library hybridization with an oligonucleotide probe. Both methods incorporate the RED (repeat expansion detection) assay that follows or identifies clones and fractions with expanded repeats. Both methods require relatively large amounts of starting DNA material and involve the use of radioactive materials for screening.

B. 2. Application of Cloning Strategy to Breast Cancer

Efficiency of the modified Glenn-Schable protocol in producing high rates of enrichment for long CAG repeats made it an optimal method to employ. This method was applied to breast cancer samples that have been identified to contain long CAG repeats using the RED assay. This was done in the framework of the previous project funded by the Department of Defence Breast Cancer Research Program (DAMD17-99-1-9302).

In order to increase the sensitivity and efficiency of the method, DNA from six RED positive breast cancer specimens were pooled and used as template for cloning. As previously described, DNA fragments obtained from constructed libraries were amplified prior to sequencing. In order to reduce the number of PCR products to be sequenced, we excluded those PCR fragments that were less than 400 bp long, assuming that such fragments are less likely to contain large repeats. The summary of colony selection and characterization is given in **Table 1**.

Among the sequenced clones using different methods, 80% of the clones contained either CAG or CTG repeats. Repeat length in the CAG containing fragments ranged from 3 to 152. Application of the modified Glenn–Schable protocol increased the percentage of CAG containing fragments from 67 % to 82% **(Table 2)**.

Table 1. Summary of colony collection, PCR-amplification and sequencing.

Task	# Total
Colonies collected, and PCR-amplified	1600
Colonies/PCR-products selected for sequencing	800
Colonies sequenced	800

Table 2. Comparison of CAG content of colonies from protocols used

CAG/CTG content	#Glenn-Schable (%)	# others (%)	# Total (%)
NO	80 (18%)	22 (33%)	102 (20%)
YES	359 (82 %)	44 (67%)	403 (80%)

Sequences obtained from all clones were blasted against the human genome using the NCBI service (http://www.ncbi.nlm.nih.gov/blast) and/or the human genome blat service at the University of California at Santa Cruz (http://genome.ucsc.edu/cgi-bin/hgBlat). The blast results were saved locally in a separate file for each sequence. The chromosomal location, number of repeats found in the genome databases and names of genes in the repeat regions were recorded. Almost all clone sequences matched the real ones in the human genome. In the case of short repeats, most of the number of repeats in our clones equaled the number of repeats in the genome database. For long repeat containing clones, the number of repeats in our clones exceeded those in the genome database.

B.3. Data Analysis and Characterization of Cloned Repeat Sequences

We have mostly finalized the sequencing of the selected 800 colonies. Data analysis at this stage is performed on less than half of the sequences. As are result of the cloning strategy we have detected that in approximately 20-30% of the repeats contained chimeric sequences from two different CAG containing sequences. These are discovered after performing blast against the human genome sequences, where the sequence matched to more than one chromosomal location. We suspect that these sequences appeared during the enrichment step of the cloning process. Secondary structures formed by repeated sequences lead to the convergence and eventual union of two fragments at the repeat region and this dramatically increases the repeat length. Chimer formation has been an insidious and limiting factor in mapping the expanded CAG repeat sequences. Chimera formation has been also reported by Inoue et al (1999) to be a limiting factor in the efficiency of cloning large TNR containing sequences.

Many CAG repeat containing clones were found to be within or in proximity to known genes **(Table 3)**. Some of these genes were listed in the NCI's Cancer Genome Anatomy Project Genetic Annotation Initiative web-site (CGAP-GAI: http://lpgws.nci.nih.gov/html-cgap/cgl/) (Clifford et al, 2000) as cancer related genes such as TBP and MLL2. CAG repeats in such genes have been reported to be polymorphic in the literature.

We have selected 116 clones containing CAG repeats (chimeras and other ambiguous sequences were excluded) and compared their lengths with those found in matching sequences from the genome databases. In the majority of the colonies (80%) there were no differences between sequence lengths of our clones and those deposited in the genome databases. Approximately 20% has shown alterations in CAG repeat lengths ranging from 1 to 10 CAG units in difference. MINK1 (Misshapen-Like Kinase 1) (Clone 0103) resembles one of the examples where the altered repeat length has observed in a CAG tract within exon 10 of this gene. We have characterized the involvement of this gene in cancer in the next paragraph. All the other sequences are studied similarly.

MINK1 (also known as B55; ZC3; MINK; hMINK; MAP4K6; MGC21111; hMINKbeta) encodes a serine/threonine kinase belonging to the germinal center kinase (GCK) family. The protein is structurally similar to kinases related to NIK and may belong to a distinct subfamily of NIK-related kinases within the GCK family. Alternative splicing occurs in this gene and four transcript variants encoding distinct isoforms have

been identified. MINK1 is found to be located at 17p13.1, which is extremely close to p53 protein on 17p13.2. As known, this chromosomal region is deleted very frequently in breast as well as other cancer types. Literature search of this gene, revealed that it has recently been identified and charachterized (Hu et al, 2004). Hu et al demonstrated that MINK1 (reported as hMNIKbeta) regulates a wide variety of cellular processes. They showed evidence of the involvement of this gene in regulating actin cytoskeleton, cell matrix adhesion, and cell-cell adhesion which leads to changes in cell morphology, cell migration and cell invasion. They also showed that the kinase inactive mutant of this gene enhanced cell-cell adhesion and blocked cell migration and cell invasion suggesting that disturbing MINK1 activity may prevent tumor progression. This study suggests that MINK1 may have an important role in cancer.

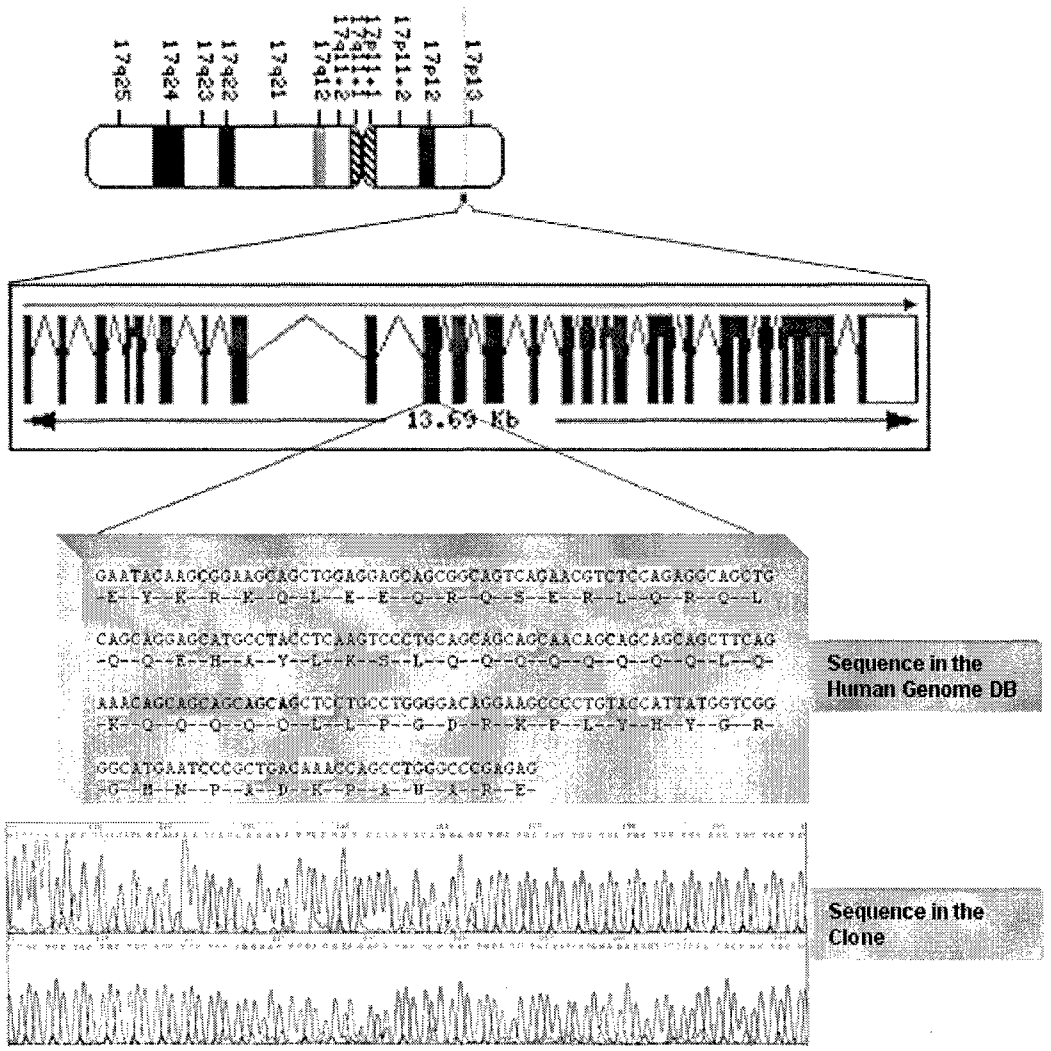


Figure 1: Map of chromosome 17 showing location of MINK 1 gene. Also shown are comparative sequences between the human genome and clone

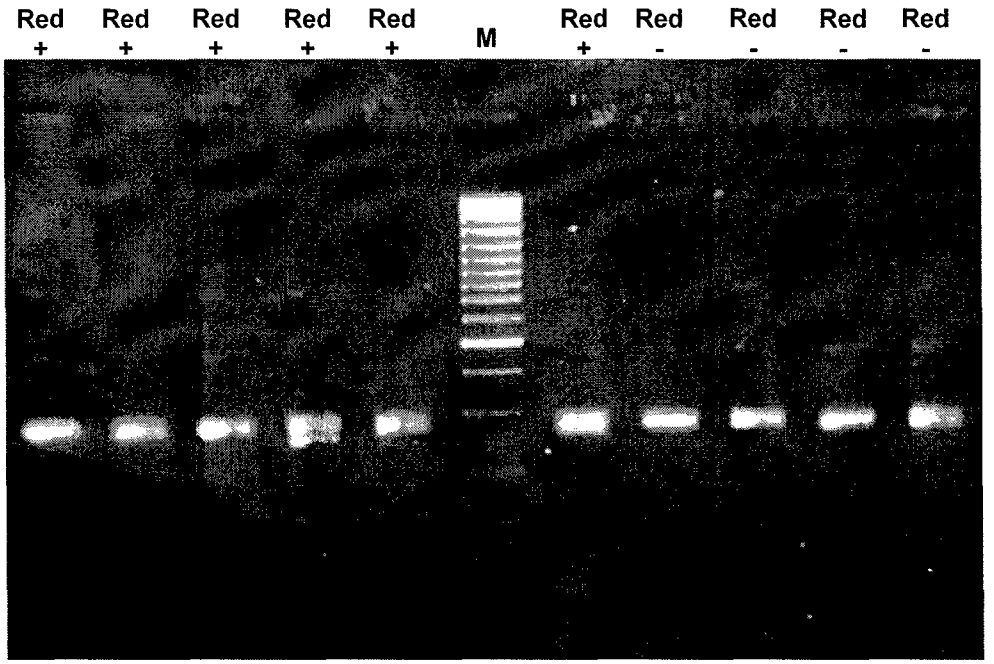
Table 3. Sequence length, repeat expansion pattern and proximity of our clones to known genes

Clone_ID	Pattern	Rep # (clone)	Chr location	Rep # (NCBI)	Nearest gene
Clone_0240	ctg	3	2	ctg(3)	ASXL2
Clone_0092	cag	10	2	(cag)10	LOC442016
Clone_0089	ctg	5	17	(ctg)5	MINK
Clone_0242	ctg	8	12	(ctg) 8	MLL2
Clone_0102	cag	21	X	21	NAP1L3
Clone_0088	cag	8	16	(cag)8	NFAT5
Clone_0301	cag	3	2	(cag)3	none
Clone_0101	tgg	3	9	(tgg)3	PAPPA
Clone_0193	ctg	5	13	(ctg)6	KIAA1704
Clone_0095	cag	10	X	(cag)11	none
Clone_0320	cag	3	12	(cag)4	none
Clone_0053	agc	4	1	(agc)3	none
Clone_0271	cag	4	20	(cag)3	none
Clone_0280	ctg	5	5	(ctg)6	none
Clone_0241	ctg	7	1	6	none
Clone_0295	cag	9	12	(cag)10	none
Clone_0096	cag	11	4	(cag)10	none
Clone_0098	cag	12	1	(cag)11	none
Clone_0286	ctg	6	3	(ctg)4	none
Clone_0293	ctg	7	3	(ctg)5	none
Clone_0296	cag	5	5	(cag)8	none
Clone_0298	cag	10	1	(cag)7	none
Clone_0200	gcct	13	2	(gcct)4	none
Clone_0103	gct	21	17	(gct)14	MINK

B.3.a. Determining the Allelic Status of the Large Repeat Regions

After identifying and localizing repeat flanking sequences in large (CAG) containing clones, we designed primer pairs for both sides of the repeats of interest using the Oligo software. We then screened the panel of patient samples that have been known to carry expanded CAG repeats (as determined by the RED method). This strategy enabled us to validate sizes of repeats between RED positive patient and control samples. In an example presented in **Figure 2**, there were no significant differences in length of PCR products between expanded and control samples. This suggests that large expansions detected by RED were not related to these loci. However, allelic change of these loci has been observed in one of the breast cancer patients. Using this procedure, we will continue to study the other candidate repeat regions suggested by the cloning method.

Figure 2. PCR products using primers designed from $(CAG)_{65}$ containing inserts RED positive and RED negative samples



B.3.b. Microsatellite Screening.

We performed microsatellite screening on detected loci that contain (CAG) repeats in their sequences using two different approaches. (i) **Screening Using ABI Microsatellite Screening System:** PCR primers were designed and optimized for amplification. PCR-amplification was performed on a panel of samples and the PCR products were sent to the microsatellite screening facility. The repeat length in each PCR product was determined and no significant variations between samples were detected for the screened loci. (ii) **Screening Using DHPLC System;** We have used this approach as a rapid way to detect polymorphism in the CAG/CTG containing loci that have been identified. We used different sets of pooled samples from breast cancer patients DNA and the other from normal population control DNA as templates for PCR. PCR products were subjected to DHPLC analysis. This method can detect any variable PCR product in pooled samples **(Figure 3).** Initial results did not show any significant differences between the samples. Other loci are still undergoing screening.

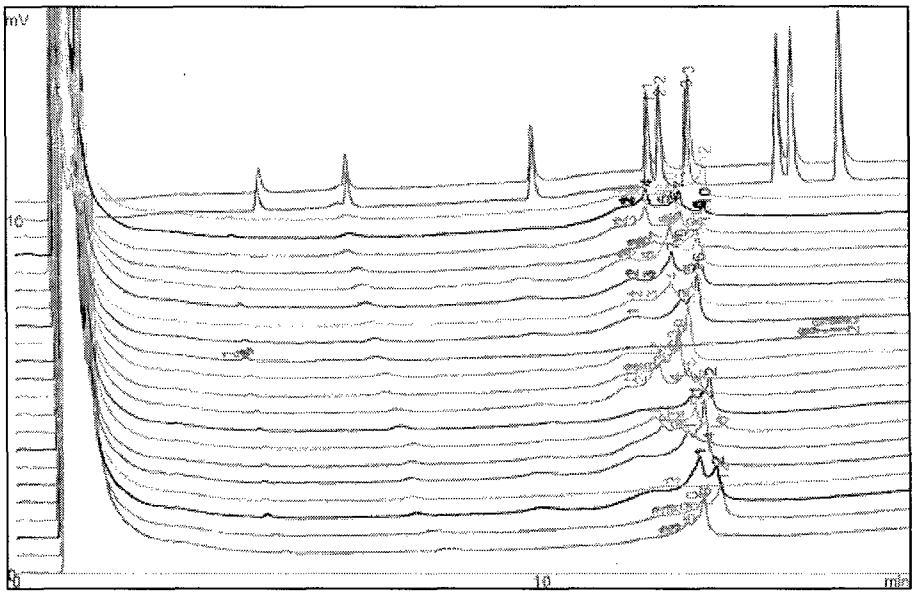


Figure 3 Chromatogram of DHPLC screening of a CAG/CTG containing locus

B.4. Systematic Mining for TNR Sequences from Cancer Genes

To fully explore the functional ramifications of the role of TNRs in cancer related genes, we have screened clones for TREs and systematically analyzed TNRs in coding regions of these genes using genomic databases and bioinformatic tools. We have systematically identified TNR sequences in coding regions of 2245 cancer related genes. We have shown that 95 TNR sequences (6 repeat and up) in 74 genes are involved in various cancer related pathways. Our results demonstrate that 57% of the repeats are located in the first exon of the genes studied. Alanine and glutamine repeats are shown to be most abundant, representing 24% and 16% of the repeats identified, respectively. Interestingly, 47% of the repeats found in known functional protein domains, suggest that repeat instability is likely to affect intrinsic properties, structure and function of proteins. Among the protein functional domains observed, signal peptide and coiled coil motifs were mostly interrupted by TR sequences. This study provides an invaluable resource for TR sequences in DNA coding regions of cancer related genes, and their potential influence on structure and function of the encoded proteins. See attached manuscript for details. Several of the repeats that have been identified in this study were also identified during the cloning application to breast cancer patients.

IV. KEY ACCOMPLISHMENTS

1. Development of a TNR cloning strategy that uses small amounts of starting DNA as a template
2. Modification of existing cloning protocol to enrich cloning of CAG containing repeats
3. Isolation of polymorphic/expanded repeat sequences in breast cancer patients and comparison to normal population controls.
4. Evaluating the contribution of polymorphic repeats within cancer related genes to breast cancer development

5. Identification of polymorphic sequences in pooled breast cancer patients (known to have expanded repeats) potentially enabling us to identify novel breast cancer predisposition genes.
6. Discovery of polymorphic CAG repeats within the coding region of MINK1 gene, which represents a potential candidate for breast cancer risk contributor.
7. Systematic analyses of CAG repeats within the coding region of ~3000 cancer genes, and characterizing the influence of repeat polymorphism on the structure and the function of protein domains
8. Comparison of molecular and bioinformatics strategies in identifying CAG repeat containing genes polymorphisms of which maybe a factor in breast cancer predisposition.

V. REPORTABLE OUTCOMES

Hamdi Jarjanazi, Hong Li, Irene L. Andrulis, and Hilmi Ozcelik, "Cloning strategy for trinucleotide repeat expansion: Searching for novel breast cancer predisposition gene(s)". Controversies in the etiology, detection and treatment of breast cancer: 2002, June 13-14, 2002 Toronto, Ontario, Canada *(Poster Presentation).*

Hamdi Jarjanazi, Hong Li, Irene L. Andrulis, and Hilmi Ozcelik, "Identification of novel breast cancer predisposition gene(s) with trinucleotide repeat expansions". The Samuel Lunenfeld Research Institute Annual Retreat: 2002, October, 9-10 YMCA Geneva Park, Ontario, Canada *(Poster Presentation).*

Hamdi Jarjanazi, Hong Li, Irene L. Andrulis, and Hilmi Ozcelik, "Identification of novel breast cancer predisposition gene(s) with trinucleotide repeat expansions". AACR 2003 Annual Meeting: 2002, April 5-9, Toronto, Ontario Canada *(Mini-symposium Presentation).*

Hamdi Jarjanazi, Hong Li, Irene L. Andrulis and Hilmi Ozcelik, "Systematic Approach to Study the Role of Repeat Sequences in Cancer-Related Genes." (Poster), Proceedings of the American Association of Cancer Research (AACR) 95th Annual meeting, March 2004, Orlando, FL, USA, (*Poster Presentation*)

Hamdi Jarjanazi, Hong Li, Irene L. Andrulis, and Hilmi Ozcelik, " Genome Wide Screening of CAG/CTG Trinucleotide Repeat Lengths in Breast Cancer" *(manuscript in preparation)*

Hamdi Jarjanazi, Noel Pabalan, Hilmi Ozcelik. "Rapid and efficient cloning of long (CAG) repeat sequences using dynabeads enrichment and DNA Array hybridization techniques". (*manuscript in preparation*)

Hamdi Jarjanazi, Noel Pabalan, Keith Wong, Hilmi Ozcelik, "Systematic Analyses of Trinucleotide Repeats in the Coding Regions of Cancer Related Genes, and their Functional Implications". (*manuscript in preparation*)

Jarjanazi et al, The work performed and the results obtained in this project will be written as a manuscript(s) and will be sent to peer review Journals with an interest in cancer genetics.

VI. CONCLUSIONS AND FUTURE WORK

We have developed an effective CAG-repeat cloning strategy. This methodology is applied to breast cancer patient specimens known to carry expanded CAG repeats. We have constructed a library by sequencing colonies with relatively large PCR products. Data analysis to date has shown several repeat sequences which show polymorphic patterns when compared with the NCBI sequences. These repeat containing sequences are highly likely to influence the intrinsic properties and function of nearby genes. Currently we are completing data analysis and bioinformatic applications to characterize genomic structures and nearby genes to repeat sequences. Isolated repeat sequences with polymorphic patterns may represent candidate breast cancer predisposition genes. After complete characterization of the repeat sequences, the best candidates will be selected and their contribution to breast cancer development will be investigated using large breast cancer and population control specimens. New funding will be sought for the accomplishment of this future objective

References

Clifford R, Edmonson MN, Hu Y, Nguyen C, Scherpbier T and Buetow KH: Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project. ***Genome Res***, 10: 1259-1265, 2000

Glenn T, Schable M. Microsatellite Isolation with Dynabeads 2003© *http://www.uga.edu/srel* 2003

Hamilton MB, Pincus EL, Di Fiore A and Fleischer RC. Universal linker and ligation procedures for construction of genomic DNA libraries enriched for microsatellites. ***Biotechniques***. Sep 27(3): 500-2, 504-7, 1999

Hu Y, Leo C, Yu S, Huang BC, Wang H, Shen M, Luo Y, Daniel-Issakani S, Payan DG and Xu X. Identification and functional characterization of a novel human misshapen/Nck interacting kinase-related kinase, hMINK beta. ***J Biol Chem***. Feb 11 280 (6): 5128, 2005

Inoue S, Takahashi K and Ohta M. Sequence analysis of genomic regions containing trinucleotide repeats isolated by a novel cloning method. ***Genomics***. Apr 157(1): 169-72, 1999

Koob MD, Benzow KA, Bird TD, Day JW, Moseley ML and Ranum LPW. Rapid cloning of expanded trinucleotide repeat sequences from genomic DNA. ***Nat Genet*** Jan18 (1): 72-5, 1998

Margolis RL, McInnis MG, Rosenblatt A and Ross CA. Trinucleotide repeat expansion and neuropsychiatric disease. ***Arch Gen Psychiatry*** 56(11): 1019-31, 1999

O'Donovan MC, Guy C, Craddock N, Bowen T, McKeon P, Macedo A, Maier W, Wildenauer D, Aschauer HN, Sorbi S, Feldman E, Mynett-Johnson L, Claffey E, Nacmias B, Valente J, Dourado A, Grassi E, Lenzinger E, Heiden AM, Moorhead

S, Harrison D, Williams J, McGuffin P and Owen MJ. Confirmation of association between expanded CAG/CTG repeats and both schizophrenia and bipolar disorder. ***Psychol Med*** 26(6): 1145-53, 1996

Pieretti M, Zhang FP, Fu YH, Warren ST, Oostra BA, Caskey CT and Nelson DL. Absence of expression of the FMR-1 gene in fragile X syndrome. ***Cell*** 66: 817-822, 1991

Sanjeeva P and Housman DE. The complex pathology of trinucleotide repeats. ***Current Opinion in Cell Biology*** 9: 364-372, 1997

Sanpei K, Takano H, Igarashi S, Sato T, Oyake M, Sasaki H, Wakisaka A, Tashiro K, Ishida Y, Ikeuchi T, Koide R, Saito M, Sato A, Tanaka T, Hanyu S, Takiyama Y, Nishizawa M, Shimizu N, Nomura Y, Segawa M, Iwabuchi K, Eguchi I, Tanaka H, Takahashi H and Tsuji S. Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. ***Nat Genet***. Nov 14(3): 277-84, 1996

Tozaki T, Inoue S, Mashima S, Ohta M, Miura N and Tomita M. Sequence analysis of trinucleotide repeat microsatellites from an enrichment library of the equine genome. ***Genome*** Apr; 43 (2): 354-65, 2000

Trottier Y, Lutz Y, Stevanin G, Imbert G, Devys D, Cancel G, Saudou F, Weber C, David G, Tora L, Agid Y, Brice A and Mandel JL. Polyglutamine expansion as a pathological epitope in Huntington's disease and four dominant cerebellar ataxias. ***Nature*** 378: 403-406, 1995

Vincent JB, Neves-Pereira ML, Paterson AD, Yamamoto E, Parikh SV, Macciardi F, Gurling HM, Potkin SG, Pato CN, Macedo A, Kovacs M, Davies M, Lieberman JA, Meltzer HY, Petronis A and Kennedy JL. An unstable trinucleotide-repeat region on chromosome 13 implicated in spinocerebellar ataxia: a common expansion locus ***Am J Hum Genet*** Mar 66(3): 819-29, 2000

Yuan QP, Lindblad-Toh K, Zander C, Burgess C, Durr A and Schalling M. A cloning strategy for identification of genes containing trinucleotide repeat expansions. ***Int J Mol Med***. Oct 8 (4): 427-31, 2001

Systematic Analyses of Trinucleotide Repeats in the Coding Regions of Cancer Related Genes, and their Functional Implications

Hamdi Jarjanazi[1,2], Noel Pabalan[1], Keith Wong[1], Hilmi Ozcelik[1,2]

[1]F. Litwin Centre for Cancer Genetics, Samuel Lunenfeld Research Institute, and Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, [2]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada

Keywords: Trinucleotide repeats, cancer genes, coding regions, bioinformatics, function, and predisposition

Running title: Trinucleotide Repeats, Protein Function and Cancer Susceptibility

Corresponding Author:

Hilmi Ozcelik, PhD

600 University Avenue, Room 992A

Toronto M5G 1X5 Ontario Canada

Tel: 416 586 4996, Fax: 416 586 8869

e-mail: Ozcelik@mshri.on.ca

Abstract

The completion of the human genome project and the availability of enormous number of bioinformatic databases and tools are valuable resources for systematic analyses of biological systems. Repetitive trinucleotide repeat (TR) sequences in the DNA coding regions are abundant in human genome and many have shown to be associated with disease susceptibility. In this study, we have systematically identified TR sequences in the coding regions of 2245 cancer related genes. We have shown 95 TR sequences (6 repeat and up) in 74 genes involved in various cancer related pathways. Our results demonstrated that 57% of the repeats are located in the first exon of the genes studied. Alanine and glutamine repeats are shown to be most abundant, representing 24% and 16% of the repeats identified, respectively. Interestingly, 47.4% of the repeats found to be located in known functional protein domains, suggesting that instability of such repeats are likely to affect the intrinsic properties, structure and the function of the proteins. Among the protein functional domains observed, signal peptide and coiled coil motifs were mostly interrupted by the TR sequences. This study provides an invaluable resource for TR sequences in the DNA coding regions of cancer related genes, and their potential influence on the structure and function of the encoded proteins.

Introduction

About 3 % of the human genome is composed of short tandem repeats (STR), also known as microsatellites or simple sequence repeats (SSRs)[1]. Such repeat sequences are highly polymorphic and are found in the form of mono, di, tri, tetra, penta and hexa- nucleotide repeats[2]. Repeat sequences occurring in the coding regions of the DNA are likely to influence the function of the encoded proteins. Among various types of microsatellites, coding regions of the genome has shown to contain significantly higher frequency of trinucleotide repeats (TRs) compared to other repeat classes[3]. This can be explained by the fact that in-frame alterations in the repeat lengths of TRs in the coding region are more likely to be tolerated than those of other repeat classes[4].

TRs have been associated with the inheritance of several genetic disorders. To date, at least 30 diseases are linked to trinucleotide repeat expansions (TRE)[5,6]. Most of these TRE loci have been associated with inherited forms of neurological and neuromuscular disorders, including Huntington disease, Fragile X syndrome, myotonic dystrophy, and Friedreich ataxia[7-9]. The mechanisms by which TRE disrupt the function of disease related proteins depend on their location in relation to the gene sequence. In different disorders, TRE were found to be located in different parts of disease-associated genes, including the intervening and coding sequences, and 5' and 3' untranslated regions. TRE in the coding region results in the alteration of the amino acid composition of the protein, thereby potentially affecting its structure and function **(Table I)**. TRE in the non-coding region may effect the regulation of transcription, and translation resulting in defective proteins[10].

In TRE associated disorders, the expansion of CAG/CTG repeats has been observed frequently **(Table I)**. In such diseases, the majority of the CAG/CTG expansions are located in the coding regions of the corresponding genes. This category of diseases is also known as polyglutamine diseases[7,11]. Recently, the expansion of alanine tracts, particularly in transcription factor genes, has been also shown to be associated with several disorders including mental retardation and malformation of the brain[12].

Microsatellite instability has been also a major focus of research in cancer. Defects in the DNA mismatch repair mechanism has shown to lead to instability of microsatellite repeats, which has been associated with multiple sporadic types of tumors, including colorectal, pancreatic, thyroid and gastric tumors[13-19]. Although TRE has been demonstrated extensively in neurodegenerative and neuromuscular disorders, it may be related to other complex genetic disorders, including cancer. The role of TR polymorphisms and TRE has not been studied extensively in cancer. In the present study, using available public databases and bioinformatic tools, we performed systematic analyses to identify TRs located in the coding regions of cancer related genes and their potential influence on the structure and function of the encoded proteins. This strategy provides candidate cancer genes, functions of which may be influenced by the polymorphism of TRs, and more likely to contribute to development of cancer.

Methods

The total of 2245 cancer related genes involved in 17 different pathways were selected from the CGAP web site (cgap.nci.nih.gov). The selection was based on genes that have transcript sequences entries in Ensembl database. The coding sequences of these genes were

downloaded from **Ensembl** genome database web site (www.ensembl.org) using **EnSamrt** Batch data/ sequence retrieval system (www.ensembl.org/EnsMart).

The TRs within the coding regions of the 2245 cancer related genes were mapped using the Unix version of Perfect Tandem Repeat Finding Program (PTRF) written by J.R. Collins and it is available online at (ncisgi.ncifcrf.gov/~collinsj/Tandem_Repeats/downloads). The PTRF input file contained coding sequences and their Ensembl transcript ID. The PTRF results were stored in an output file that contained the Ensembl transcript ID's, repeat pattern, unit length, number of repeats and start and end position of each repeat. For further systematic analyses of the repeats, a local relational database was created and contained the following tables: 1) Ensembl data, which contains gene names, Ensembl transcript ID, Locus link ID, Ensembl peptide ID, GO ID and gene names. 2) CGAP data, which contains gene name, biological pathway classification and description of each gene. 3) PTRF results, which contains Ensembl transcript ID, repeat patterns, unit length, number of repeats, repeat start position, repeat stop position and repeat ID.

Using the above tables, a summary table containing repeat data, gene name and pathway information was generated. For TRs with repeat length of six and more, the exact exonic location of the repeat and the coded amino acid sequence was determined and recorded using the transcript structure data from Ensembl database. Using the peptide structure data in the same database, the functional domains where the TR occurs was determined for each repeat and added to the summary table. We also searched the protein data bank (PDB) at http://pdbbeta.rcsb.org/ for the availability of 3D structure for the proteins harboring the TR and the PDB ID was recorded for proteins with available 3D structures. Finally, we performed a

comprehensive Pubmed literature search for each TR to determine whether they have been studied in the context of their functional effects on the protein and contribution to disease susceptibility.

Results

We have studied the coding regions of 2245 CGAP genes (cgap.nci.nih.gov) involved in 17 different biological cancer-related pathways for the presence of coding DNA TRs, and identified 83,401 of them ranging from 2 to 27 in repeat length **(Figure 1)**. As expected, TRs with repeat length of two units were predominant compared to larger repeats. Because small size repeats exist in large numbers in the genome, we have only focused on TRs that have length of at least 6 units (n=95) for further analysis. For each coding DNA TR, we have also determined the corresponding number of amino acid repeats in the protein sequences. This has shown that the length of amino acid repeats in the corresponding protein sequence may vary when compared to nucleotide repeats in the DNA (**Tables II & V**). **Table II** shows the distribution and the types of amino acid residues (cysteine, lysine, threonine, proline, aspartic acid, glycine, histidine, glutamic acid, serine, leucine, glutamine, and alanine) coded by 95 coding DNA TRs (6 and up) found in cancer related genes. Among all, alanine was found to be the most abundant amino acid repeat (n=23, 24%) with a range of 6 to 20 amino acid repeats in length, followed by glutamine (n=15, 16%) with a range of 6 to 38 amino acid repeats in length. Interestingly, we have shown that 56 (59%) of the 95 repeats were found to be located in exon 1, whereas the rest were found to be distributed between exon 2 to 39 of the corresponding genes (**Table III**). Interestingly, among immunology genes, 83.3% of the repeats were found to be in exon 1, and 66.7% of all the repeats were alanine. We have also shown that 45 (47.4%) of the repeats occur in a known protein domain determined by Ensembl (pfam, print or pfscan) (**Tables IV & V**).

Signal peptide motif (n=10) of the proteins studied was found to be the most frequently interrupted functional motif in this study, followed by coiled coil motifs (n=8) (**Table IV**). The 95 TR sequences were observed in genes involved in various cancer pathways including immunology (33), development (18), transcription (16), tumor suppressor/oncogenes (7), signal transduction (5), pharmacology (3), cell signaling (1), DNA damage (1), metastasis (1) and miscellaneous (10) (**Table V**). Comprehensive PubMed search revealed that only 20 (20%) of the repeats were previously screened for their polymorphic status and/or repeat expansion at least in one study.

Discussion

The completion of the human genome project and the availability of enormous number of bioinformatic data and tools have provided valuable resources for systematic analyses of biological systems and their variations. Despite their abundance, repetitive sequences and their biological function, are among the least studied and understood feature of the human genome. Many genetic disorders have been associated with variations in the repetitive sequences. Particularly, neurodegenerative and neuromuscular diseases have been associated with TR sequences presenting in the genes of the related disorders. These findings suggest that the TR alterations may be also related to other complex genetic disorders, yet to be studied.

In this study, we have systematically analyzed the abundance of TR sequences in the coding regions of 2245 cancer related genes, and shown the presence of 95 TRs in 74 genes with a DNA repeat length of 6 to 27. The corresponding amino acid sequences were ranged from 4 to 38. The change in the range is due to the location of the repeat within the open reading frame of the protein and the composition of the flanking sequences of the repeats. It is however

interesting to note that the majority of the DNA repeat lengths corresponds to larger amino acid repeat lengths. Therefore, searching for repeat sequences in protein sequences may reflect a more accurate representation of the repeat lengths in DNA coding regions.

As shown in various disorders, TRs are susceptible to alterations in length, which may impact on the intrinsic properties or the function of the mRNA molecules and their encoded proteins. RNA molecules can form secondary structure elements which may affect many cellular processes including mRNA stability, transcription, RNA processing and translation. Functionally important RNA secondary structures can be found in both coding and non-coding regions of genes. Most of RNA secondary structure studies focused on 3' and 5' UTR regions and only few studies examined the RNA secondary structures in coding regions [20]. Changes in TR lengths in the coding regions might affect the secondary structure of mRNA coded by such sequences. In a recent review, Yeap and coworkers showed that the CAG TR in exon 1 of the Androgen Receptor (AR) gene forms a stable stem-loop structure in which increasing repeat number increases the length of the stem preserving the stem-loop motif[21]. The study showed that changes in the stem-loop length affect the AR mRNA stability and its interaction with RNA binding proteins.

Expansion of TR sequences within the encoded proteins is well studied in polyglutamine disorders. While the mechanisms remain unknown, many studies have shown that the expanded poly-glutamine tracks affect protein stability and solubility and leads to formation of nuclear inclusions that might contribute to disease[22,23]. A study on the polyglutamine repeat in the ataxin-2 gene, contributing to spinocerebellar ataxia type 2 (SCA2), showed that the expansion of the glutamine track alters the Golgi localization, causing cell death[24]. Coding TRs

of AR are very well studied examples, indicating the impact of TRs on cancer development. AR contains two polymorphic repeats in its N-terminal; a poly-glutamine (CAG)n at codon 58 and a poly-glycine (GGC)n at codon 442. The CAG repeat length ranges from 14-25 whereas the GGC repeats range from 16-18 in healthy individuals[25]. Studies have demonstrated that shorter CAG and CCG repeats in AR are associated with prostate cancer risk[25-29]. Another recent study has shown that the presence of one or two long (CAG)n alleles or a cumulative (CAG)n repeat size exceeding 42 may be associated with a slight increase in breast cancer risk whereas the presence of one or two long (GGC)n alleles may be associated with a substantive reduction in breast cancer risk in young women[30].

In cancer related genes studied, alanine and glutamine repeats were found to be more abundant and relatively larger in repeat length compared to other amino acid repeats. Alanine (6-20 repeat length) and glutamine (6-38 repeat length) tracks represented 24% and 16% of all repeat sequences, respectively (**Table II**). The polymorphic length of glutamine repeats associated with the disease genes ranged between 4 and 44 in normal population, and 36 to 306 in the disease[8,31]. Similarly, alanine repeats ranged between 4-18 amino acids in normal, and 5-29 amino acids in disease population[12]. These results demonstrate that the ranges of glutamine and alanine repeat lengths in our data set are comparable to those associated with disease. This suggests that such repeats in cancer related genes are susceptible to alterations and/or expansion through various instability mechanisms, and may represent good candidates for disease predisposition. None of the other coding repeats (cysteine, lysine, threonine, proline, aspartic acid, glycine, histidine, glutamic acid, serine, and leucine) were shown to be associated with any disease to date, but the fact that some are quite large can suggest that they may be unstable, and thus may play a role in disease susceptibility. The predominance of alanine and

glutamine repeats over other types found in this study is consistent with previous findings by other researchers. Karlin and coworkers reported that alanine, glutamine, proline and glycine repeats are more abundant in human genes than other amino acid repeats. They suggested that high alanine frequency in proteins may reflect on alpha-helix stability and its flexible hydrophobic properties[32]. In a screen of 80,000 proteins from the Swiss-Prot Database, Katti and coworkers found that 14% of proteins contain significant internal repeats and that among the proteins containing 10 or more amino acid repeats, glutamine, alanine, glycine, glutamic acid, and serine repeats were much more frequent than other amino acid repeats[33]. They suggested that long tandem repeats of highly hydrophobic amino acids are probably not favored in proteins and hydrophilic amino acids, could be tolerated to a considerable extent if they occur in the linker regions and if they can be easily solvated on surface of the protein.

Interestingly, we have observed that 59% of all the coding repeats are located in the first exon of the cancer genes studied. Among 15 genes in the immunology category, 83% of the repeats were in exon 1, and 67% of them consisted of alanine repeats. This interesting observation suggests a mechanism of action for alanine repeats located in exon 1 of the immunology genes studied. Huntley and coworkers have found that developmental proteins are also enriched with alanine repeats, speculating that such repeats allow for protein elongation and functional specialization of the repeat region via mutation[34]. Alanine repeats are hydrophobic, therefore, they tend to occupy internal positions in the protein[35].

Theoretically, alteration of the repeat length of poly-amino acid tracks in proteins would influence their folding patterns and in turn, their native 3-dimensional structure. Interestingly, we have shown that 45 (47%) of the 95 repeats occurred in a known protein domain determined by

Ensembl (pfam, print or pfscan) (**Table IV & V**). These repeats, in polymorphic state, are likely to alter the properties of such functional units, which may result in the alterations of the intrinsic properties and the function of the proteins. Unfortunately, to date X-ray structure has not yet been resolved for any polymorphic poly-glutamine track containing protein, hence the effect of repeat length alterations on the 3-dimensional structure of the polyglutamine containing proteins remains to be elucidated[36].

In our study, 10 out of 45 (22%) of the repeat sequences were found to be located within signal peptide motifs and 9 of them (90%) consisted of leucine tracks. These repeats are likely to alter the signal sequences, affecting the translocation of the genes within the cell. Leucine repeats are common in signal peptide sequences near the amino terminus of membrane and extracellular proteins. Unlike other aliphatic and aromatic residues in the human genome, leucine runs are more common. The prominence of leucine among protein sequences reflects its important role in hydrophobic cores in transmembrane segments and in signal peptides, and its prevalence and stability in secondary and tertiary structures. Transmembrane segments of receptors or extracellular proteins are usually enriched with non-specific hydrophobic runs[32]. Our data obtained through screening of the CGAP cancer related genes are in accordance with these finding. We have found that 11 proteins (12%) contained leucine repeats and 10 of them occurred within a signal peptide or transmembrane domain.

The second most common functional domain interrupted by repeat sequences was coiled coil domains (18%). Coiled coil motifs are found in wide variety of proteins with distinct function and structures and they play role protein oligomerization and in protein-protein interaction. For example, they are found in transcriptional factors, in structural fibre networks in muscle and the

cytoskeleton, in the molecular machineries that mediate mitosis, and in certain membrane associations. Coiled coil regions in proteins range in length from tens to hundreds of residues[37,38].

Conclusion

In the light of the above examples, TRs in coding regions play an important role in understanding the molecular basis of genetic disorders. Among the 95 repeats analyzed in this study, only 20 (20%) have been studied and all shown to be polymorphic. This supports the fact that TRs are susceptible to changes in size length. It is also evident from this search that the significance of repeat sequences on the protein function or disease association have not been studied extensively. Our analyses showed that 47.4% of these repeats are located in a known functional protein domain, which suggest that the alterations in repeat lengths may disturb the function of such proteins. Thus, our systematic analysis of TRs in cancer related genes provides an invaluable resource for investigators interested in the function and cancer association of genetic alterations.

Acknowledgements

We would like to thank to Mehjabeen Shariff for bioinformatics assistance, and Dr. Sevtap Savas for critical revision of the manuscript. This work was supported by projects (DAMD17-99-1-9302 and DAMD17-02-1-0499) from Department of Defense Breast Cancer Research Program, which is managed by the U.S. Army Medical Research and Materiel Command.

References

1. Weber JL, May PE. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 1989;44(3):388-396.

2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM,

Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. Initial sequencing and analysis of the human genome. Nature 2001;409(6822):860-921.

3. Subramanian S, Madgula VM, George R, Mishra RK, Pandit MW, Kumar CS, Singh L. Triplet repeats in human genome: distribution and their association with genes and other genomic regions. Bioinformatics 2003;19(5):549-552.

4. Jasinska A, Krzyzosiak WJ. Repetitive sequences that shape the human transcriptome. FEBS Lett 2004;567(1):136-141.

5. Pearson CE. Slipping while sleeping? Trinucleotide repeat expansions in germ cells. Trends Mol Med 2003;9(11):490-495.

6. Yuan QP, Lindblad-Toh K, Zander C, Burgess C, Durr A, Schalling M. A cloning strategy for identification of genes containing trinucleotide repeat expansions. Int J Mol Med 2001;8(4):427-431.

7. Usdin K, Grabczyk E. DNA repeat expansions and human disease. Cell Mol Life Sci 2000;57(6):914-931.

8. Vincent JB, Paterson AD, Strong E, Petronis A, Kennedy JL. The unstable trinucleotide repeat story of major psychosis. Am J Med Genet 2000;97(1):77-97.

9. Ohara K. Anticipation, imprinting, trinucleotide repeat expansions and psychoses. Prog Neuropsychopharmacol Biol Psychiatry 2001;25(1):167-192.

10. Sinden RR, Potaman VN, Oussatcheva EA, Pearson CE, Lyubchenko YL, Shlyakhtenko LS. Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. J Biosci 2002;27(1 Suppl 1):53-65.

11. Margolis RL, McInnis MG, Rosenblatt A, Ross CA. Trinucleotide repeat expansion and neuropsychiatric disease. Arch Gen Psychiatry 1999;56(11):1019-1031.

12. Brown LY, Brown SA. Alanine tracts: the expanding story of human illness and trinucleotide repeats. Trends Genet 2004;20(1):51-58.

13. Gryfe R, Gallinger S. Microsatellite instability, mismatch repair deficiency, and colorectal cancer. Surgery 2001;130(1):17-20.

14. Karran P. Microsatellite instability and DNA mismatch repair in human cancer. Semin Cancer Biol 1996;7(1):15-24.

15. Onda M, Nakamura I, Suzuki S, Takenoshita S, Brogren CH, Stampanoni S, Li D, Rampino N. Microsatellite instability in thyroid cancer: hot spots, clinicopathological implications, and prognostic significance. Clin Cancer Res 2001;7(11):3444-3449.

16. Palli D, Russo A, Ottini L, Masala G, Saieva C, Amorosi A, Cama A, D'Amico C, Falchetti M, Palmirotta R, Decarli A, Costantini RM, Fraumeni JF, Jr. Red meat, family history, and increased risk of gastric cancer with microsatellite instability. Cancer Res 2001;61(14):5415-5419.

17. Shimizu Y, Ikeda S, Fujimori M, Kodama S, Nakahara M, Okajima M, Asahara T. Frequent alterations in the Wnt signaling pathway in colorectal cancer with microsatellite instability. Genes Chromosomes Cancer 2002;33(1):73-81.

18. Speicher MR. Microsatellite instability in human cancer. Oncol Res 1995;7(6):267-275.

19. Yamamoto H, Itoh F, Nakamura H, Fukushima H, Sasaki S, Perucho M, Imai K. Genetic and clinical features of human pancreatic ductal adenocarcinomas with widespread microsatellite instability. Cancer Res 2001;61(7):3139-3144.

20. Katz L, Burge CB. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res 2003;13(9):2042-2051.

21. Yeap BB, Wilce JA, Leedman PJ. The androgen receptor mRNA. Bioessays 2004;26(6):672-682.

22. Davies SW, Beardsall K, Turmaine M, DiFiglia M, Aronin N, Bates GP. Are neuronal intranuclear inclusions the common neuropathology of triplet-repeat disorders with polyglutamine-repeat expansions? Lancet 1998;351(9096):131-133.

23. Reid SJ, van Roon-Mom WM, Wood PC, Rees MI, Owen MJ, Faull RL, Dragunow M, Snell RG. TBP, a polyglutamine tract containing protein, accumulates in Alzheimer's disease. Brain Res Mol Brain Res 2004;125(1-2):120-128.

24. Huynh DP, Yang HT, Vakharia H, Nguyen D, Pulst SM. Expansion of the polyQ repeat in ataxin-2 alters its Golgi localization, disrupts the Golgi complex and causes cell death. Hum Mol Genet 2003;12(13):1485-1496.

25. Ding D, Xu L, Menon M, Reddy GP, Barrack ER. Effect of GGC (glycine) repeat length polymorphism in the human androgen receptor on androgen action. Prostate 2004b.

26. Beilin J, Ball EM, Favaloro JM, Zajac JD. Effect of the androgen receptor CAG repeat polymorphism on transcriptional activity: specificity in prostate and non-prostate cell lines. J Mol Endocrinol 2000;25(1):85-96.

27. Buchanan G, Irvine RA, Coetzee GA, Tilley WD. Contribution of the androgen receptor to prostate cancer predisposition and progression. Cancer Metastasis Rev 2001;20(3-4):207-223.

28. Giovannucci E, Stampfer MJ, Krithivas K, Brown M, Dahl D, Brufsky A, Talcott J, Hennekens CH, Kantoff PW. The CAG repeat within the androgen receptor gene and its relationship to prostate cancer. Proc Natl Acad Sci U S A 1997;94(7):3320-3323.

29. Hsing AW, Gao YT, Wu G, Wang X, Deng J, Chen YL, Sesterhenn IA, Mostofi FK, Benichou J, Chang C. Polymorphic CAG and GGN repeat lengths in the androgen receptor gene and prostate cancer risk: a population-based case-control study in China. Cancer Res 2000;60(18):5111-5116.

30. Suter NM, Malone KE, Daling JR, Doody DR, Ostrander EA. Androgen receptor (CAG)n and (GGC)n polymorphisms and breast cancer risk in a population-based case-control study of young women. Cancer Epidemiol Biomarkers Prev 2003;12(2):127-135.

31. Cummings CJ, Zoghbi HY. Fourteen and counting: unraveling trinucleotide repeat diseases. Hum Mol Genet 2000;9(6):909-916.

32. Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. Amino acid runs in eukaryotic proteomes and disease associations. Proc Natl Acad Sci U S A 2002;99(1):333-338.

33. Katti MV, Sami-Subbu R, Ranjekar PK, Gupta VS. Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. Protein Sci 2000;9(6):1203-1209.

34. Huntley MA, Golding GB. Neurological proteins are not enriched for repetitive sequences. Genetics 2004;166(3):1141-1154.

35. Perutz MF, Pope BJ, Owen D, Wanker EE, Scherzinger E. Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of Sup35 and of the amyloid beta-peptide of amyloid plaques. Proc Natl Acad Sci U S A 2002;99(8):5596-5600.

36. Ding D, Xu L, Menon M, Reddy GP, Barrack ER. Effect of a short CAG (glutamine) repeat on human androgen receptor function. Prostate 2004a;58(1):23-32.

37. Hicks MR, Walshaw J, Woolfson DN. Investigating the tolerance of coiled-coil peptides to nonheptad sequence inserts. J Struct Biol 2002;137(1-2):73-81.

38. Mason JM, Arndt KM. Coiled coil domains: stability, specificity, and biological implications. Chembiochem 2004;5(2):170-176.

39. Buchanan G, Yang M, Cheong A, Harris JM, Irvine RA, Lambert PF, Moore NL, Raynor M, Neufing PJ, Coetzee GA, Tilley WD. Structural and functional consequences of glutamine tract variation in the androgen receptor. Hum Mol Genet 2004;13(16):1677-1692.

40. Nakamura K, Jeong SY, Uchihara T, Anno M, Nagashima K, Nagashima T, Ikeda S, Tsuji S, Kanazawa I. SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. Hum Mol Genet 2001;10(14):1441-1448.

41. Utsch B, Becker K, Brock D, Lentze MJ, Bidlingmaier F, Ludwig M. A novel stable polyalanine [poly(A)] expansion in the HOXA13 gene associated with hand-foot-genital syndrome: proper function of poly(A)-harbouring transcription factors depends on a critical repeat length? Hum Genet 2002;110(5):488-494.

Table I. Trinucleotide repeats in coding regions of genes associated with human diseases

Gene and Disease	Protein	Repeat pattern	Repeat size (wt)	Repeat size (disease)	Refs
Coding Region (Poly Glutamine Diseases)					
HD	Huntingitin	CAG	6-35	36-121	[31]
SBMA	AR Exon 1	CAG	6-39	40-62	[39]
DRPLA/HRS	Atrophin1	CAG	6-35	49-88	[31]
SCA1	Ataxin-1	CAG	6-44	39-82	[31]
SCA2	Ataxin-2	CAG	15-31	36-63	[31]
SCA3 (MJD1)	Ataxin-3	CAG	12-40	55-84	[31]
SCA7	Ataxin-7	CAG	4-35	37-306	[31]
SCA17	TBP	CAG	29-42	47-55	[40]
Coding region (Poly Alanine Diseases)					
HOXA13, Hand-foot-genital syndrome (HFGS)	Homeodomain protein, transcription factor	GCG	18	24-26	[41]
HOXD13, Synpolydactyly (SPD)	Homeodomain protein, transcription factor	GCG	15	22-29	[41]
CBFA1(RUNX2,) Cleidocranial dysplasia (CCD) *NX2)* Cleidocranial	Homeodomain protein, transcription factor	GCG	17	27	[41]
Zic2, Holoprosencephaly (HPE)	Zinc finger protein, transcription factor	GCG	15	25	[41]
PABP2, Oculopharyngeal muscular dystrophy(OPMD)	Poly(A) binding protein	GCG	10	11-17	[41]
GPX1	Gluatthione peroxidase	GCG	4	5-6	[41]

Figure 1: Distribution of DNA repeat number of coding trinucleotide repeats in cancer genes

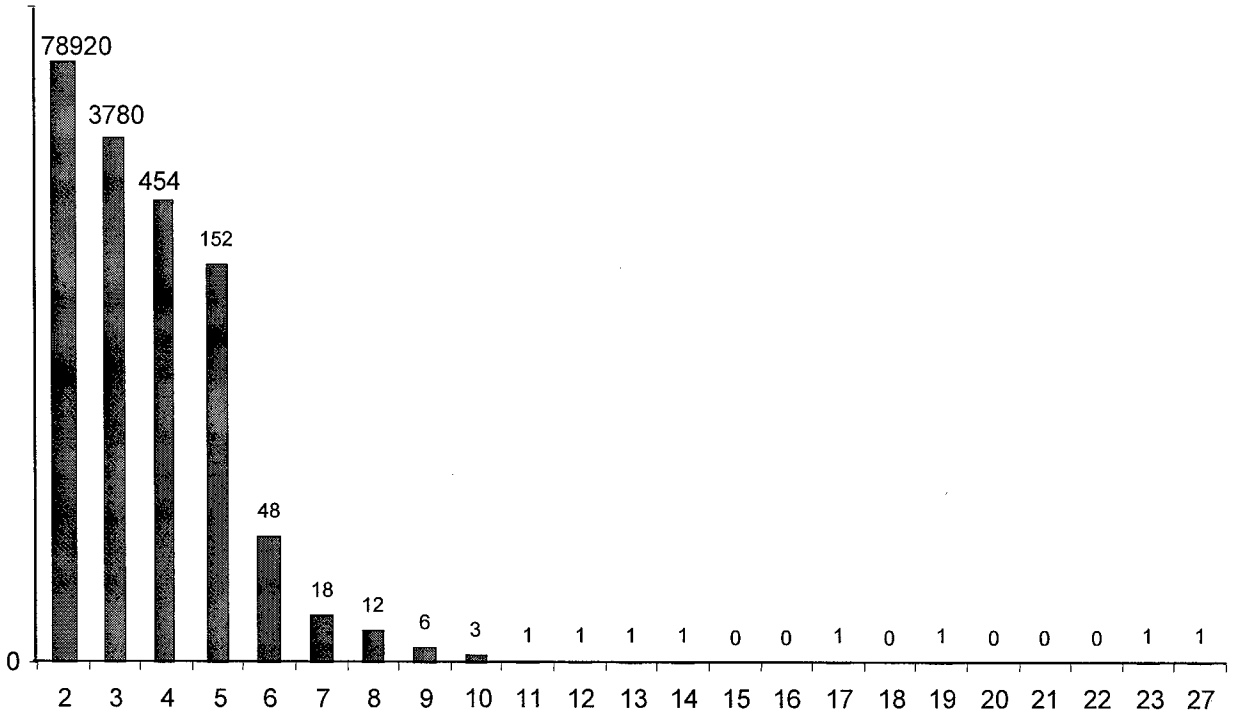


Table II. Number of trinucleotide repeats loci coding for repetitive different amino acids.

Number of TR (n=95)	Repeated aa length range	Amino acid (poly)	Amino Acid coded	Associated with disease
1	6	C	Cysteine	No
1	6	K	Lysine	No
1	7	T	Threonine	No
4	7-11	P	Proline	No
6	5-8	D	Aspartic acid	No
7	8-23	G	Glycine	No
7	6-12	H	Histidine	No
9	4-27	E	Glutamic acid	No
10	5-11	S	Serine	No
11	6-10	L	Leucine	No
15	6-38	Q	Glutamine*	Yes
23	6-20	A	Alanine*	Yes

Table III: Cancer pathways and the properties of the coding DNA trinucleotide repeats identified

Pathways	Repeats # (%)	Exon 1 # (%)	Functional Domains # (%)	A content # (%)	Q content
Immunology	18	**15 (83.3)**	9 (50)	**12(66.7)**	1
Development	33	20 (60.6)	17(51.5)	6(18.2)	6
Transcription	16	7 (43.8)	6(37.5)	2(12.5)	4
Others	28	14 (50%)	13(46.4)	3(10.7)	4

Table IV: Functional domains containing polymorphic trinucleotides in cancer genes

Functional Protein Domains (n=45)	No of repeats	Frequency (%)	Function of the motif & cancer relevance
Signal peptide	10	22.2	Signal peptides indicate a protein that will be secreted
Coiled Coil	8	17.8	Found in a wide variety of proteins. Structurally they are composed of two or three alpha helices that wind around each other
Proline-rich	4	9	They either structural proteins or metal-binding proteins. They can also bind DNA.
AR	3	7	It has 3 functional and structural domains: an N-terminal (modulatory) domain; a DNA binding domain and a hormone binding domain.
NLS_BP	2	4	A frequent hit producer. In the absence of other evidence , a match to this entry should only be taken as a weak indication of nuclear localization.
Granin	2	4	Acidic proteins present in the secretory granules of endocrine and neuro-endocrine cells. Granins may be precursors of biologically-active peptides, or they may be helper proteins in the packaging of peptide hormones and neuropeptides - their precise role is unclear.
Antifreeze	2	4	Type I AFPs are Ala-rich, amphiphilic, alpha-helical proteins
Others	14	31	-

Table V: Trinucleotide repeats in cancer related genes

Gene	Pathway	DNA Repeat	DNA Repeat #	Amino Acid	Amino Acid #	Exon	Protein Domains
ADRA2B	Immunology	agg	6	E	9	2	Rhodopsin like G-protein-coupled receptors (GPCR)
ALPP	Immunology	gct	7	L	7	1	Signal peptide
APOB	Immunology	ctg	6	L	6	1	Transmembrane / Signal peptide
APP	Immunology	cac	7	T	7	6	-
AR	Immunology	gcg	17	G	23	1	-
AR	Immunology	cag	6	Q	6	1	Androgen receptor
AR	Immunology	cag	23	Q	23	1	Androgen receptor / Coiled Coil
ASCL1	Development	cag	12	A	13	1	Coiled Coil
ATBF1	Transcription	aac	10	E	15	8	Coiled Coil
ATBF1	Transcription	gcg	8	G	14	9	-
ATBF1	Transcription	agc	7	Q	8	10	-
ATRN	Immunology	cgg	6	A	7	1	-
ATRN	Immunology	ctg	6	L	8	1	-
AXIN2	Development	cca	6	H	6	5	-
BLMH	Pharmacology	cgc	9	P	9	1	Proline rich
BMP6	Development	gca	7	Q	6	1	Transforming growth factor beta (TGFb), N-terminal
CBL	Tumor Suppressor/Oncogenes	acc	7	H	7	1	-
CBL	Tumor Suppressor/Oncogenes	tga	6	D	5	9	-
CBX4	Immunology	acc	11	H	12	3	-
CENPB	Immunology	gag	6	E	27	1	Coiled Coil
CHD4	Transcription	gga	7	E	6	3	-
CHD4	Transcription	aga	6	K	6	3	Bipartite nuclear localization signal (NLS_BP)
CHES1	Tumor Suppressor/Oncogenes	tcc	6	S	6	7	-
CHGA	Immunology	agg	8	E	9	6	Granin
CHGA	Immunology	agg	6	E	8	7	Granin
CTNND2	Signal transduction	cgc	6	A	6	7	-
CTNND2	Signal transduction	gcc	8	P	7	7	Proline rich
DDX10	Immunology	gat	8	D	8	17	-
EGR1	Transcription	agc	6	S	6	1	-
EGR2	Immunology	ccg	6	A	10	2	-
EN1	Development	gcg	6	A	6	1	Antifreeze 1
EOMES	Development	ccg	7	A	14	1	Antifreeze 1
EPHB6	Signal transduction	ctc	8	S	11	4	Ephrin receptor
FGF10	Miscellaneous	ctg	6	C	6	1	Transmembrane /Signal peptide
FGFR1	Signal transduction	atg	6	D	7	5	-
FUS	Miscellaneous	gcg	6	G	10	6	-

GAS6	Immunology	ctg	6	L	6	1	signal peptide
GATA6	Miscellaneous	cca	10	H	10	1	GATA-type transcription activator, N-terminal (GATA-N)
GSPT1	Immunology	gcg	10	G	11	1	-
HD	Immunology	cgc	7	P	11	1	Proline rich
HD	Immunology	agc	19	Q	21	1	Coiled Coil
HGFAC	Miscellaneous	tgc	6	L	10	1	Transmembrane / signal peptide
HLXB9	Development	ccg	9	A	14	1	Orphan nuclear receptor, NOR1 type
HLXB9	Development	gcg	6	G	8	1	Eggshell
HOXA1	Development	cac	8	H	8	1	-
HOXA13	Development	ccg	6	A	18	1	-
HOXA13	Development	ccg	6	A	6	2	-
HOXA2	Development	ccg	6	A	6	1	-
HOXD11	Development	cgg	6	A	13	1	-
HOXD8	Immunology	cgg	6	A	9	1	-
ID4	Transcription	cgg	6	A	10	1	-
IL16	Immunology	ctc	6	S	6	1	-
IL9R	Immunology	gca	8	S	8	9	-
IRS1	Immunology	gca	7	Q	6	1	-
IRS1	Immunology	agc	7	S	7	5' utr, 1	-
ITGAL	Metastasis	ctg	6	L	6	30	Integrins alpha chain / Transmembrane
JUND	Tumor Suppressor/Oncogenes	ccg	7	A	10	1	JNK MAP kinase
KCNN3	Pharmacology	gca	8	Q	12	1	Proline rich
KCNN3	Pharmacology	agc	14	Q	14	1	-
LAF4	Transcription	cag	7	S	7	12	AF-4 proto-oncoprotein
MAF	Tumor Suppressor/Oncogenes	gcg	8	G	14	1	Maf transcription factor (TF Maf)
MAF	Tumor Suppressor/Oncogenes	cca	6	H	6	1	-
MLL2	Miscellaneous	agc	7	Q	7	39	Coiled Coil
MLL2	Miscellaneous	cag	7	Q	11	39	Coiled Coil
MLLT1	Miscellaneous	ctc	6	S	5	6	-
MLLT4	Miscellaneous	agg	6	E	4	28	-
MOG	Immunology	cct	7	L	6	1	Transmembrane/Signal peptide
MSH4	DNA damage	gca	6	S	6	1	-
NOTCH4	Tumor Suppressor/Oncogenes	tgc	9	L	10	1	Signal peptide
NPM1	Signal transduction	atg	6	D	6	6	Nucleoplasmin
NUCB1	Immunology	gca	7	Q	7	11	-
PAPPA	Immunology	gcc	7	A	7	2	-
POU3F1	Transcription	cgg	8	A	11	1	-
POU3F1	Transcription	acc	6	H	6	1	-
POU3F2	Transcription	ggc	6	G	21	1	-
POU3F2	Transcription	agc	6	Q	21	1	Coiled Coil

PPP3CA	cell signaling	cgg	9	A	9	1	-
PVRL1	Immunology	agg	8	E	8	6	-
PVRL2	Immunology	ctg	6	L	7	1	Signal peptide
PVRL2	Immunology	agg	6	E	5	9	-
QSCN6	Miscellaneous	ctg	6	L	6	1	Signal peptide
RXRB	Immunology	cgg	6	A	7	1	-
SATB1	Transcription	agc	7	Q	15	11	-
SMARCA2	Transcription	agc	13	Q	24	3	Coiled Coil
SMARCC2	Development	ctc	6	P	7	27	Proline rich extensin signature
SOX1	Development	gcg	6	A	8	1	-
SOX1	Development	gcg	8	A	9	1	-
TBP	Immunology	agc	27	Q	38	2	CC
TGFBR1	Development	gcg	9	A	9	1	Activin receptor / transmembrane
TNFAIP2	Immunology	gcg	6	A	6	2	-
TOP2B	Transcription	gat	6	D	6	31	-
TSC1	Miscellaneous	cag	6	S	6	23	-
WNT6	Development	ctg	6	L	7	1	Transmembrane / Signal peptide
ZIC3	Development	gcc	9	A	9	1	-
ZNF6	Transcription	atg	6	D	6	4	Zfx / Zfy transcription activation region